Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 204–216
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Expression recognition using fuzzy spatio-temporal modeling

T. Xiang, M.K.H. Leung*, S.Y. Cho

School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore

Received 7 October 2005; received in revised form 16 October 2006; accepted 27 April 2007

Abstract

In human–computer interaction, there is a need for computer to recognize human facial expression accurately. This paper proposes a novel and effective approach for facial expression recognition that analyzes a sequence of images (displaying one expression) instead of just one image (which captures the snapshot of an emotion). Fourier transform is employed to extract features to represent an expression. The representation is further processed using the fuzzy C means computation to generate a spatio-temporal model for each expression type. Unknown input expressions are matched to the models using the Hausdorff distance to compute dissimilarity values for classification. The proposed technique has been tested with the CMU expression database, generating superior results as compared to other approaches.

Keywords: Facial expression; Fourier transform; Fuzzy C means; HCI; Hausdorff distance; Spatio-temporal

1. Introduction

During the past two decades, facial expression recognition (FER) has attracted a significant interest in the scientific community for its many applications, such as emotion analysis, interactive video, indexing and retrieval of image and video databases, image understanding, and synthetic face animation [1]. Among them, the most important potential application is human computer interaction [2]. It is crucial for computers to be able to interact with the users, in a way similar to human-to-human interaction [3]. Psychological studies have indicated that humans interact with each other mainly through speech, but also through display of emotions for emphasis purpose [4]. One of the important way humans display emotions is through facial expressions. There are six facial expressions associated universally with six distinct emotions as smile, sadness, surprise, fear, anger and disgust [5,6]. Most facial expressions recognition systems have been examined using these six basic expressions.

Till now numerous approaches have been proposed and promising results have been reported. However, some performed their experiments in a relatively controlled environment because their systems have limitations such as

1. they require many feature points and landmarks to be detected inside the facial area [2,4,7];
2. they are sensitive to the variations of face scale [8], lighting condition [9,10], and facial component shape [11–13];
3. they require a predefined physical facial model [4,14].

In this paper, we propose a novel fuzzy spatio-temporal (FST) approach for real time FER. The proposed system first employs the Fourier transform to convert a facial expression sequence of images (displaying one expression) from the spatio-temporal domain into the spatio-frequency domain. This is followed by a fuzzy C means (FCM) [15] classification for expression representation. Six expression models can then be constructed. Unknown input expressions are matched to the models using the Hausdorff distance to compute dissimilarity values for classification. The goal of the work presented here is to develop a recognition system that can tackle or tolerate some of the above limitations to certain extent with the following focuses on design.

1. *Dynamic information analysis*: A sequence of images (displaying one expression) instead of just one image (which

* Corresponding author.
E-mail address: asmkleung@ntu.edu.sg (M.K.H. Leung).

0031-3203/$30.00
doi:10.1016/j.patcog.2007.04.021

captures the snapshot of an emotion) is analyzed. This is based on the psychological study in Ref. [16] that temporal information can reveals more relevant information about the underlying emotional states than a single static image. For example, it is difficult to distinguish a static image of a slightly smiling face from a face with neutral expression. On the other hand, the expression can be easily perceived if the smiling face image is displayed right after the corresponding neutral face image.

2. *Expression descriptor search*: In this study, we favor and search for a local descriptor for expression representation. A global pattern descriptor is not appropriate since it cannot be extended to the occluded and cluttered environment. The search needs to tackle the following issues:
   (a) *Dimension*: A sequence of images spans a three-dimensional space of position (x–y) and time (t). A local descriptor can be built on top of the one-dimensional intensity changes with respect to t at each pixel position.
   (b) *Robustness*: To cater for lighting changes, Fourier transform is employed in the description to transform intensity expression into frequency information that is less influenced by lighting change.
   (c) *Effectiveness/efficiency*: It is likely that more information can lead to better result if the information is relevant. On the other hand, if redundant or irrelevant information is included, the result can deteriorate and take much longer time to compute. Hence, it is of interest here to find/experiment with the right amount of information to obtain optimal or near optimal result that can be computed efficiently.
3. *Recognition mechanism search*: FCM [15] is one useful approach for fuzzy classification, which can determine the intrinsic division in a set of unlabeled data and find representatives for homogeneous groups. Unlike hard C-means algorithm which executes a sharp classification in which each object is either assigned to a class or not, FCM can make an object belong to several classes at the same time but with different degrees between 0 and 1 [17]. Since some expressions might look alike at certain facial regions, it is hard to put clear boundaries between different expressions. Thus, FCM is used since it can give intermediate memberships instead of hard boundaries. To tolerate misalignment and variations of facial component shape, a fast, non-one-to-one matching technique, the Hausdorff distance [16], is added to aid the matching process.
4. *Ease of automation*: Unlike other systems that need many facial landmarks, the proposed system needs only the positions of two eyes that can be detected accurately in current technology. Since fewer landmarks are needed, the system is less affected by the inaccuracy of facial features detection. Consequently, the system can be automated easily in future.
5. *Expression model construction*: A training approach is used to generate expression models automatically. No effort is needed to predefine physical facial model.

The proposed technique has been tested with the CMU expression database, generating superior results as compared to other approaches.

The rest of the paper is organized as follows. Section 2 gives a short review of FER methods. Following this, the proposed approach is described in Section 3 in details. Section 4 shows the experimental results, using the Cohn–Kanade expression database [30] as the test data. Discussion and conclusion are drawn in Section 5.

2. Literature review

FER has been extensively studied in recent years [9,14,18,19]. Due to different ages, genders and external objects such as glasses and hair, automatic FER is a challenging task. Furthermore, faces appear different because of changes in pose and lighting condition. Variations such as these should be considered and handled by an automatic FER system [20]. A wide range of algorithms has been applied to the automatic FER problem [10,21–23]. The two main types can be described as spatio-temporal and spatial approaches [24].

2.1. Spatio-temporal approaches

In Ref. [7], Ekman and Friesen built perhaps the best known system for classifying facial expressions based on what they call action units (AU). Each AU corresponds to several muscles that together form certain facial action. From then on, numerous improved and ameliorated AU systems have been proposed [13]. In the early 1990s, the engineering community started to use these results to construct automatic methods for recognizing emotions from facial expressions in images or video [11]. Tian and Kanade improved the AU recognition algorithm by using a more efficient facial features tracking system [12]. They reported their system was more automatic and more robust than previous models since in their system, the extracted features were represented and normalized based on an explicit face model that was invariant to image scale and head motion. Lien et al. explored HMM for facial AU recognition in Ref. [25]. Each AU or AU combination was assigned a specific HMM topology according to the pattern of feature motions. A directed link between states of the HMM represented the possible inherent transition from one facial state to another. An AU was identified when its associated HMM had the highest probability among all HMMs. However, since each AU or AU combination is associated with one HMM, the approach is infeasible for covering a great number of potential AU combinations involved in facial expressions. Furthermore, all AU systems suffer a common disadvantage that the combination of different AUs results in a large number of possible facial expressions, which makes the problem extremely complex. A system based on HMM and using MPEG-4 parameters for analysis was reported in Ref. [5]. A video sequence was first extracted and then analyzed by a semi-continuous HMM. In Ref. [26], Oliver, Pentland, and Berard also used HMM to classify the facial expression. Cohen et al. employed Bayesian network classifier for FER by using the maximum likelihood estimation [27]. The system described in

Ref. [10] detected face regions and extracted facial features in four direction: horizontal, vertical, and diagonally in both directions. A robot started learning facial expression by recognizing these features and actions. In Ref. [28], Essa and Pentland developed an automated system using optical flow coupled with geometric, physical and motion-based dynamic models of face muscles to describe facial motions. Facial expression identification was based on the invariance between the motion energy template learned from ideal two-dimensional motion views and the motion energy of the observed image. Since their system needed a physical face model, the classification accuracy therefore relied much on the validity of such a model. The method is also limited as it can only identify typical facial action. In practice, there are many variations of a typical facial action; each showing a different pattern.

2.2. Spatial approaches

Neural network (NN) systems are often used in FER [29]. They are applied either directly on face images or combined with principal component analysis (PCA), independent component analysis (ICA) and Gabor wavelets filter [8,30]. In Ref. [31], Avent et al. developed a low-level system to detect the edge clusters corresponding to the eyes, eyebrows, and lips. The edges were exploited in the classification process using NN. Lisetti and Rumelhart [32] proposed another NN-based approach to recognize the facial expressions using different facial regions (manually cropped). In Ref. [33], Feitosa et al. applied two different models of NN: back-propagation and RBF network as classifiers of facial expression. In Ref. [34], Franco and Treves built a NN which structure was implemented with specialized modules, and trained with back-propagation. Zheng et al. [35] manually located 34 landmark points from each facial image and then converted these geometric points into a labeled graph (LG) vector using the Gabor wavelet transformation method to represent the facial features. For each training facial image, the semantic ratings describing the basic expressions were combined into a six-dimensional semantic expression vector. Analysis of the correlation between the LG vector and the semantic expression vector was performed by the Kernel Canonical Correlation Analysis (KCCA). According to this correlation, the associated semantic expression vector of a given test image was estimated. In Ref. [18], a face representation based on appearance models was used and two bilinear factorization models were subsequently proposed to separate expression and identity factors from the global appearance parameters. The authors concluded that bilinear factorization could outperform these techniques in terms of correct recognition rates and synthesis photorealism especially when the number of training samples was restricted. In Ref. [36], Guo and Dyer introduced a new technique based on linear programming for both feature selection and classifier training in the FER system. In Ref. [37], fuzzy integrals were used to describe the uncertainty of facial expression. Facial expression space could be constructed automatically and compared for expression classification. Most of the above work suffered from a common limitation that only static clue from still face images was used for FER. No temporal behaviors were taken into consideration, which might reveal more underlying facial expression information.

3. The proposed approach

The architecture of the proposed facial expression modeling and recognition system is displayed in Fig. 1. The input is a sequence of images displaying one particular facial expression, changing from a neutral state with no emotion to a state with extreme muscle distortion. Each image will be preprocessed for eye positions detection and facial area localization, such that a face can be detected and normalized. For each pixel inside the facial area, its intensity value will change over time when there is muscle movement at this location due to expression. The changes form a unique temporal pattern. Each pattern will be modeled and represented using Fourier transform.

- When the input is the model data for training, it will go through the FCM extraction process [17] to find suitable numbers of classes to classify each pixel-based temporal pattern. Finally, each expression type is modeled by an expression structure which is the combination of the pixel-based FCM classification results and the geometrical locations of pixels with respect to the eyes.
- When the input is the test data of an unknown facial expression, the pre-determined FCM is used to classify the temporal pattern of each pixel. The classification results are then combined with the pixel locations to form an expression structure which will be matched to six model expression structures to classify the input expression.

In the following, each of the processes will be described in detail.

3.1. Preprocessing

Each image will be preprocessed to compute the eye positions, such that the facial area can be extracted and normalized. Currently, the eye positions have been extracted manually in order to test system optimal performance with respect to the FST concept. Let the number of frames in an expression sequence be M. M cannot be set low here since the recognition rate will drop due to information loss. Currently, it is set as 11. Let the left eye position at frame k be $(EyeX(k), EyeY(k))$ with $1 \leqslant k \leqslant M$. According to the position of the left eye, a fixed-size facial area can be identified for analysis as shown in Fig. 2. The area should be normalized (scaled) by D/d_k, where D and d_k are the standard and actual distances between eyes. Since the downloaded database had been normalized with a fixed distance of 110 for both D and d_k, the normalization step has been skipped.

3.2. Pixel-based temporal pattern extraction

Within the facial area, the intensity value at each pixel position can be represented as $f(x, y, t)$, where x and y specify the location while t is the timing parameter with $1 \leqslant t \leqslant M$. Each $f(x, y, t)$ over time should form a unique temporal pattern and is modeled and represented using the Discrete Fourier

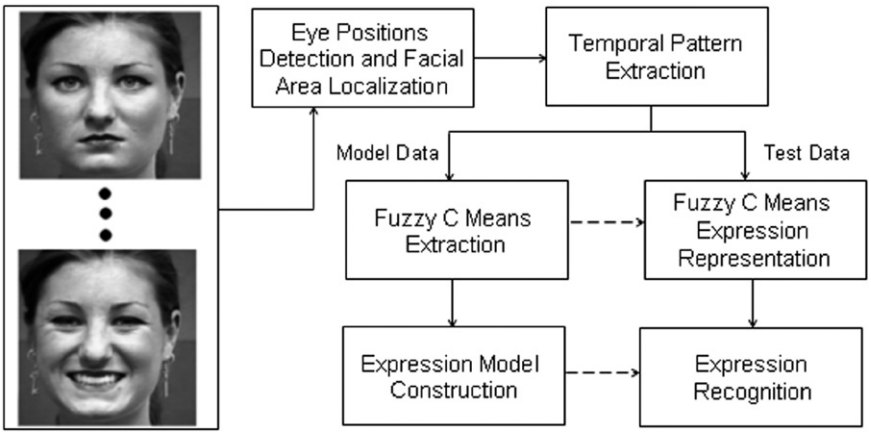

Fig. 1. The architecture of the facial expression modeling and recognizing system.

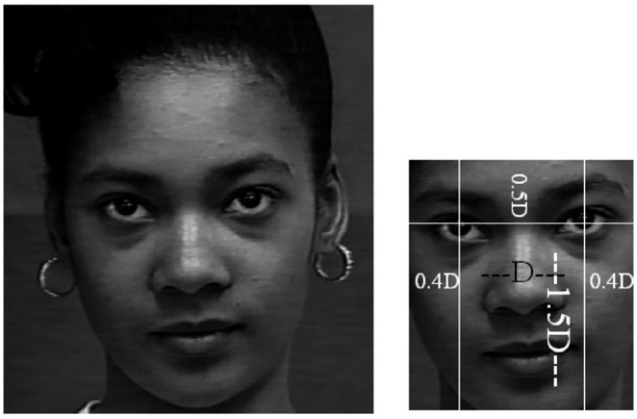

Fig. 2. An example of facial area detection where D is the reference distance between two eyes.

transform [38] as

$$F(x, y, m) = \frac{1}{M}\sum_{k=1}^{M} f(x, y, t)\mathrm{e}^{-\mathrm{j}t2\pi m/M} \quad (1)$$

with $m = 1, 2, \ldots, M$, and "j" represents imaginary number. Since $\mathrm{e}^{-\mathrm{j}t2\pi m/M} = \cos(2\pi tm/M) + \mathrm{j}\sin(2\pi tm/M)$, Eq. (1) can be rewritten as

$$F(x, y, m) = \frac{1}{M}\sum_{t=1}^{M} (f(x, y, t)\cos(2\pi tm/M) + \mathrm{j}f(x, y, t)\sin(2\pi tm/M))$$

with two parts: the real part

$$real(x, y, m) = \frac{1}{M}\sum_{t=1}^{M} f(x, y, t)\cos(2\pi tm/M)$$

and the imaginary part

$$img(x, y, m) = \frac{1}{M}\sum_{k=1}^{M} f(x, y, t)\sin(2\pi tm/M).$$

The first component ($F(x, y, 1)$) which reflects the lighting intensity level will be ignored while the other components which capture the frequency information of the temporal pattern will be kept to form representation that is relatively insensitive to lighting change. In addition, it is relatively rare for a temporal pattern to have high frequency (Fourier) components because the input captures one expression only and the time period to perform such expression is short. With these considerations, we collect the first N_d Fourier component pairs (the real and imaginary parts), with $m = 2, 3, \ldots, N_d + 1$, to represent and describe a pixel-based temporal pattern as

$$v(x, y) = [real(x, y, 2) \quad img(x, y, 2) \quad real(x, y, 3) \quad img(x, y, 3) \quad \ldots \quad real(x, y, N_d + 1) \quad img(x, y, N_d + 1)]^{\mathrm{T}}.$$

Currently, N_d is set as 4, and the average information loss according to the following formula is only 12.1%. This 12.1% information can include frequencies from noise that we do not want to process:

$$Loss = \sum_{m=6}^{\lfloor M/2\rfloor+1} |F(x, y, m)| \Big/ \sum_{m=2}^{\lfloor M/2\rfloor+1} |F(x, y, m)| .$$

Pixel selection: To speed up computation, one can consider those pixels that exhibit significant expression energy only, i.e. each pixel is ranked by

$$Sum(x, y) = \sum_{m=2}^{N_d+1} |real(x, y, m)| + |img(x, y, m)|. \quad (2)$$

Fig. 3. Examples of different types of mouth movements with expressions (a) neutral, (b) smile, (c) surprise and (d) sad.

We have experimented with different percentages of the top ranking pixels and found that 20% gave the best performance (see Section 4.2). At this moment, let the number and set of the selected pixels be $N_{\%}$ and $S_{\%}$.

3.3. Expression model construction and training

An expression can be represented and described by its specific muscle movements at different locations inside the facial area. For example, the mouth movement is quite dramatic for smile and surprise. Comparatively, it is much mild for the sad expression. Fig. 3 shows four examples of mouth movements for four expressions. Consequently, the modeling needs to record both the location and movement information. Section 3.3.1 models the movement patterns while the location information will be combined in Section 3.3.2.

When the inputs are the model expression sequences for training, the inputs will go through the "Pixel-based fuzzy C means extraction" and the "Expression model construction" processes (see Fig. 1). The movement pattern at each position is modeled in the first process. We have modeled the movement at each pixel using $f(x, y, t)$, and subsequently $F(x, y, m)$ and $v(x, y)$. The feature vector $v(x, y)$ is further refined using the FCM method to find c classes. For each expression, we assume that the number of pixel-based movement patterns is limited. It is of interest to find such a number, c, and then classify the patterns into c different classes accordingly. Finally, each of the six expressions (smile, surprise, anger, disgust, fear and sad) is modeled by the membership function of the FCM. Details of the construction are described below.

3.3.1. Pixel-based FCM extraction

It is observed that the pixel-based temporal pattern, $f(x, y, t)$, exhibits unique patterns of one peak, two peaks or a mixtures of them as shown in Fig. 4 where three examples of $f(x, y, t)$ and its $v(x, y)$ are shown. These patterns can be extracted and modeled using the FCM method [17] to classify each input vector $v(x, y)$ according to c fuzzy subclasses by minimizing the following sum of square errors:

$$J(K_i, n) = \sum_{j=1}^{n} \sum_{k=1}^{c} (u_{ijk})^m d^2(v_{ij}, \mu_{ik}). \tag{3}$$

In Eq. (3), the unknowns are u_{ijk} and μ_{ik}. The meaning of each symbol is

- i: This is referring to the ith expression training, i.e. one of the six mentioned expressions.
- K_i: The vector set for training the ith expression model.
- n: The number of $v(x, y)$ in K_i. It is equal to $N_i \times N_{\%}$.
- N_i: The number of training sequences for expression i.

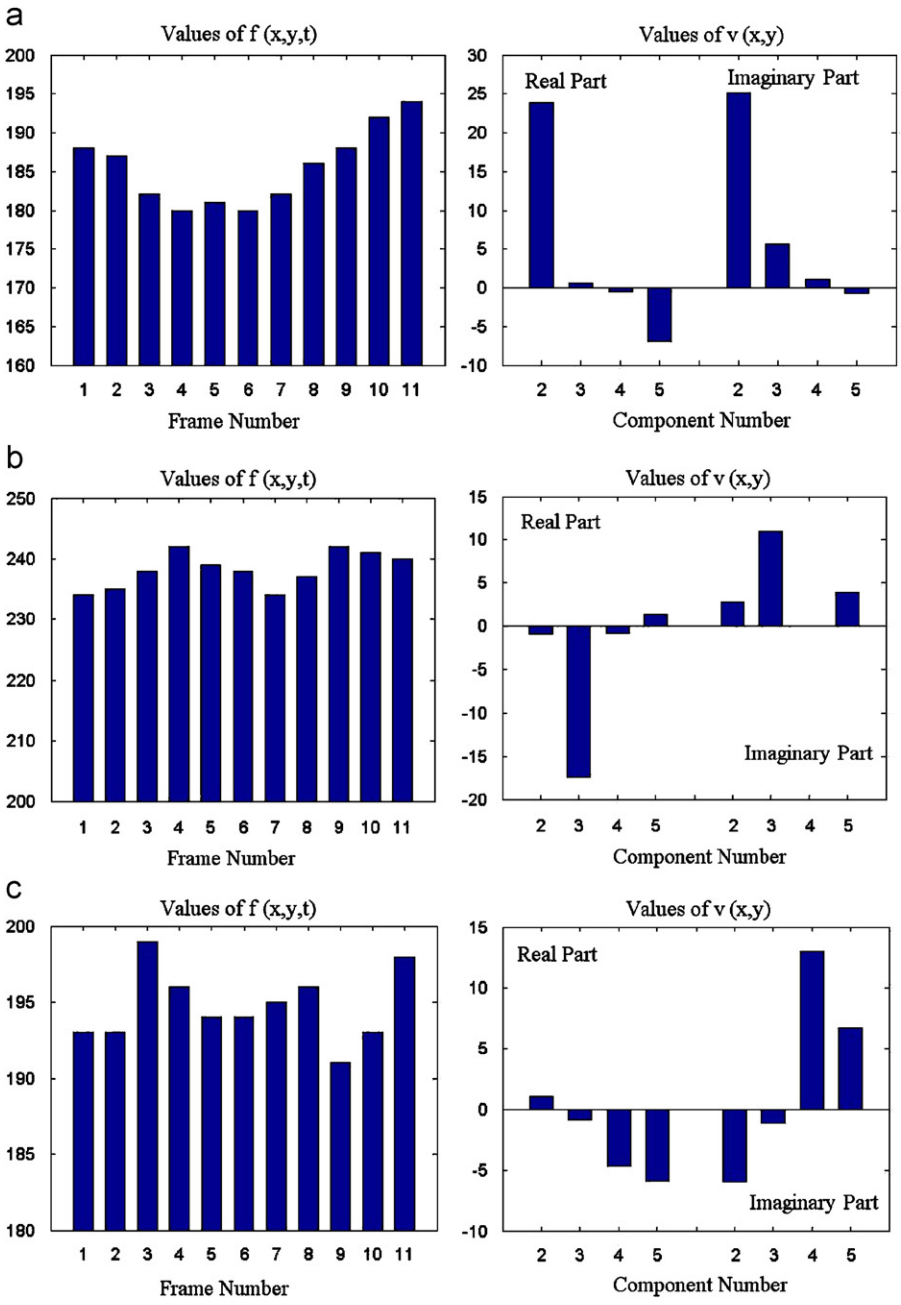

Fig. 4. Examples of $f(x, y, t)$ and $v(x, y)$ with (a) one peak; (b) two peaks; (c) mixture of peaks.

- c: The number of fuzzy subclasses. Experiments will be performed to find the optimal c.
- v_{ij}: The short hand form for $v_{ij}(x, y)$.
- u_{ijk}: The short hand form for $u_{ijk}(x, y)$ which is the degree of fuzzy membership of input vector $v_{ij}(x, y)$ being in the kth subclass.
- m: The fuzzy exponent ($m \in [1, \infty]$), which determines the degree of overlapping between classification subclasses. If m is close to or equal 1, FCM becomes a hard k means clustering. On the other hand, the boundaries of sample cluster membership become increasingly blurred if m is set too high. According to Ref. [17], we set m to 2.
- μ_{ik}: The kth subclass center (or mean vector) of the ith expression.
- $d^2(v_{ij}, \mu_{ik})$: The distance (Euclidian distance is used here) from v_{ij} to μ_{ik}.

In Eq. (3), u_{ijk} should satisfy the following conditions:

$$u_{ijk} \in [0, 1] \quad \forall i, j, k,$$
$$\sum_{k=1}^{c} u_{ijk} = 1,$$
$$0 \leqslant \sum_{j=1}^{n} u_{ijk} \leqslant n \quad \forall i, k. \tag{4}$$

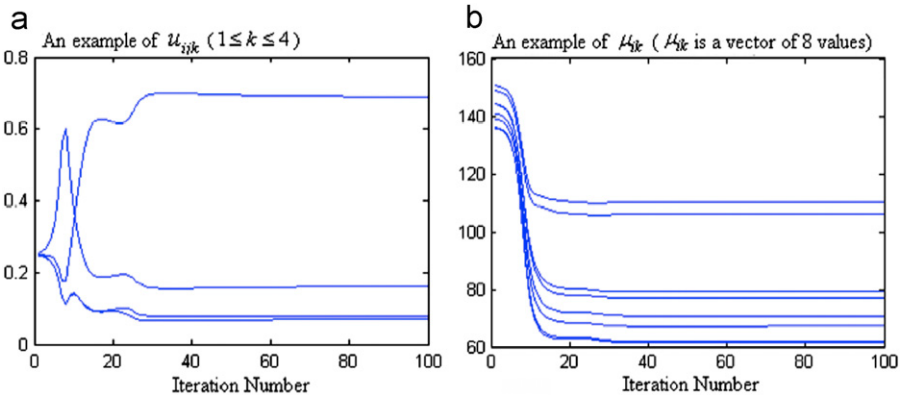

Fig. 5. Typical examples of the convergence of u_{ijk} (a) and μ_{ik} (b).

According to Ref. [17], the minimization of Eq. (3) provides an iterative solution of u_{ijk} and μ_{ik} as

$$u_{ijk} = \frac{1}{\sum_{p=1}^{c} (d(v_{ij}, \mu_{ik})/d(v_{ij}, \mu_{ip}))^{2/(m-1)}}, \quad (5)$$

$$\mu_{ik} = \frac{\sum_{j=1}^{n} u_{ijk}^{m} v_{ij}}{\sum_{j=1}^{n} u_{ijk}^{m}}. \quad (6)$$

The iterative algorithm is outlined in the following:

Initialization: $u_{ijk} = u_{ijk}^{0} = 1/c$ (Assume equal probability initially.)
$z = 1$. (z is a counter)
Repeat the following:
Begin
Update u_{ijk} and μ_{ik} using Eqs. (5) and (6)
If error is small, i.e. $\max_{ijk}\{|u_{ijk}^{z+1} - u_{ijk}^{z}|\} < \varepsilon$, then stop iteration.
If exceed maximum number of iterations, i.e. $z > max\_iteration$, then stop iteration.
$z = z + 1$
End repeat

The values of *max_iteration* and ε are set as 100 and 10^{-5}, respectively. In practice, the computation converges in about 60 iterations. Typical examples of the convergences of u_{ijk} and μ_{ik} over time are shown in Fig. 5. The initial value of u_{ijk} is set as 0.25 since $c = 4$ gives the best experimental results (see Section 4.2).

3.3.2. Expression model construction

Each expression model, $Expr_i$, composes of two sets, the motion subclasses $\{\mu_{ik}$ with $1 \leqslant k \leqslant c\}$ and the membership function matrix $\{u_{ik}(x, y)$ with $1 \leqslant k \leqslant c\}$ as

$$Expr_i = \{\{\mu_{ik} \text{ with } 1 \leqslant k \leqslant c\}, \{u_{ik}(x, y) \text{ with } 1 \leqslant k \leqslant c\}\}. \quad (7)$$

Each entry of the matrix $u_{ik}(x, y)$ specifies the fuzzy membership value of subclass k for each expression. $u_{ik}(x, y)$ is related to the generated $u_{ijk}(x, y)$ from previous section. Since $u_{ijk}(x, y)$ is computed from the top 20% pixels (Eq. (2)) of significant motion from each sequence, only 20% of the facial area from each sequence has valid member function values while the other areas get a value of zero. The average values of $u_{ijk}(x, y)$ over N_i sequences will then give $u_{ik}(x, y)$ as

$$u_{ik}(x, y) = \frac{1}{N_i} \sum_{p=x, q=y} u_{ijk}(p, q), \quad (8)$$

where p and q are determined by j (see Eq. (3)). Fig. 6 shows examples of $u_{ik}(x, y)$ with $1 \leqslant i \leqslant 6$ and $1 \leqslant k \leqslant 4$. Each image has been normalized by the maximum value of all the u_{ik}. These model membership matrices will be used in the recognition process to match against the input membership matrices to classify an input expression.

3.4. Facial expression recognition

When the input is a test sequence for expression recognition or classification, the sequence will go through the "Fuzzy C means expression representation" process and the "Expression recognition" process in Fig. 1. The movement pattern at each pixel will be classified in the first process according to the pre-determined FCM, $\{\mu_{ik}$ with $1 \leqslant i \leqslant 6$, $1 \leqslant k \leqslant c\}$. Consequently, the input fuzzy membership matrix $\{w_{ik}(x, y)$ with $1 \leqslant i \leqslant 6$, $1 \leqslant k \leqslant c\}$ can be computed using Eq. (5). In the expression recognition process, we compute the dissimilarity value (DV_i) between the input $w_{ik}(x, y)$ and the models $u_{ik}(x, y)$ as

$$DV_i = \sum_{x=1}^{col} \sum_{y=1}^{row} D_i(x, y), \quad (9)$$

$$D_i(x, y) = \min_{(m,n) \in R(x,y)} \sum_{k=1}^{c} |w_{ik}(x, y) - u_{ik}(m, n)|, \quad (10)$$

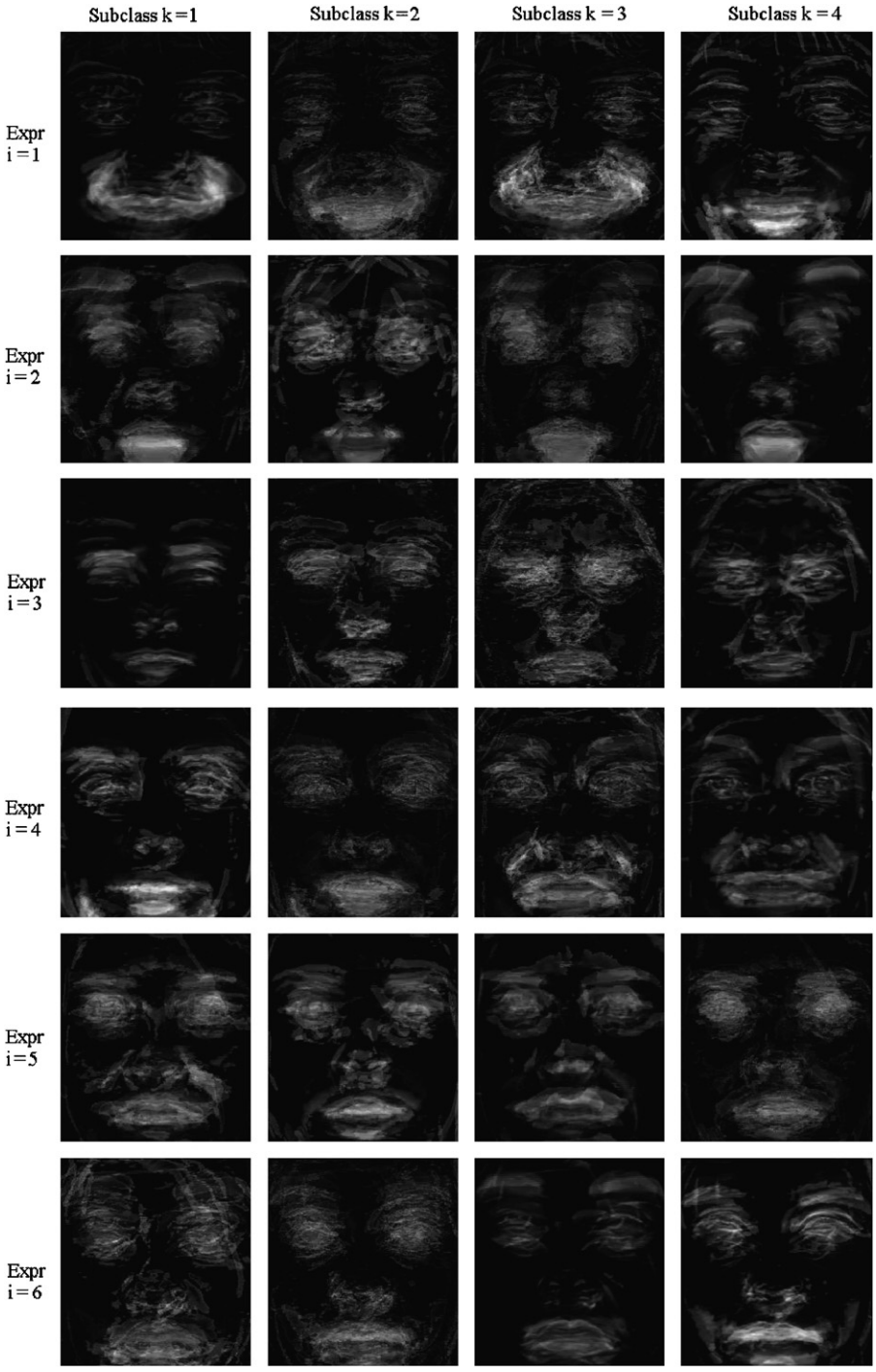

Fig. 6. Membership matrices u_{ik} of the six basic expressions.

where *row* and *col* are the total numbers of rows and columns of pixels inside the facial area. $R(x, y)$ is a small neighborhood centered at (x, y) with size of $(col/40) \times (row/40)$.

Refinement with respect to movement extent: It can be observed that different expressions have different extent of muscle movement. We measure the movement extent (using Eq. (2)) of each sequence, k, as

$$ME(k) = \sum_{(x,y)\in S_{\%}} Sum(x, y).$$

The means and variances of *ME* can be computed as

$$\mu_i = \frac{1}{N_i}\sum_{k=1}^{N_i} ME(k), \quad \sigma_i^2 = \frac{1}{N_i - 1}\sum_{k=1}^{N_i} (ME(k) - \mu_i)^2.$$

With an unknown expression sequence, the probability of it belonging to expression i can be computed as

$$P_i = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(\frac{-(ME - \mu_i)^2}{2\sigma_i^2}\right).$$

With P_i, we can refined DV_i as

$$DV_i' = DV_i * (1 - P_i). \quad (11)$$

Finally, the ith expression match that produces the minimal dissimilarity, DV_i', is selected as the match.

4. Experimental results

To evaluate performance of the proposed algorithm, we have run experiments on the popular Cohn–Kanade facial expression database [30]. It consists of 96 subjects, 391 expression sequences with resolution of 640×480 and frame rate at 30 frames per second. It captures full-face frontal views under constant illumination at relatively the same scale of facial area. Each expression recording starts at neutral expression (no expression) and ends at the peak of the muscle distortion. Many other systems had evaluated their performances on this database, and this facilitates comparison of our system to other expression recognition systems. In Cohn–Kanade database, each subject plays one or more than one of the basic six expressions, but not necessarily all the six expressions. In addition, some sequences have less than 11 frames that exhibit movements. Hence, we cannot use all 96 subjects in our experiment. A summary of the numbers of subjects and expression sequences that were used in our experiment is tabulated in Table 1.

Excluding the background, each facial area has a resolution of 200 × 220. The frame numbers are different for different sequences in the Cohn–Kanade database. To include as many sequences as possible in our experiments, we accept sequences that have at least 11 frames of expression activity. Sequences with lesser number of frames could have too little information to yield meaningful interpretation, and hence they will be excluded. If a sequence has more than 11 frames of expression activity, some frames with lesser activity will be eliminated by the following procedure:

1. Assume that there are K frames in the original sequence but M frames should be kept. If $K < M$, this sequence will be ignored; else if $M = K$, this sequence is accepted. In either case, we will not execute steps 2–6.
2. In current setting, we have $M = 11$ and will keep the first and the last frames.
3. Let K' be the number of frames in current sequence. If $K' = M$, it is done.
4. For each tth $(2 \leqslant t \leqslant K' - 2)$ frame, calculate the difference picture as

$$Difference(t) = \sum_i \sum_j |f(i, j, t+1) - f(i, j, t)|.$$

5. Select the minimal $Difference(t)$ and delete the corresponding tth frame.
6. Repeat steps 3–6.

Table 1
Numbers of facial expression sequences of the Cohn–Kanade database

Number of subjects	Smile	Surprise	Anger	Disgust	Fear	Sad
95	93	88	25	33	54	72

With the correct number of frames per sequence, the proposed system was run according to the following parameters listed in Table 2.

4.1. Expression recognition

Experiments on FER can be conducted in two different manners: subject dependent (part of the data for each subject is used as training data while the other part is used as test) and subject independent (data of the test subject is not used as training data). Cohen had done experiments using both approaches [4] while some [11,29] conducted only the subject dependent test. We chose to do the subject independent test for the following reasons:

- Each subject played at most once for each expression type in the Cohn–Kanade database. Hence, it was not possible to carry out the dependent test on this database.
- Independent test happens more frequently in real applications and it is more difficult. In Ref. [4], the authors reported recognition rates of 83.31% and 65.11% for the dependent and independent test, respectively. Similarly, the best dependent recognition rate in Ref. [39] was 97.5%, much higher than the best independent recognition rates (86.4%) in the same experiments. In Ref. [35], the authors tested their system on two different databases. The best dependent test results were 98.4% and 81.3%, while the best independent test results were 77.1% and 79.2%. All dependent tests have shown better performance than independent tests.

Our experiment was conducted in the following manner. We used the data of all but one subject as training data, and tested on the sequences of the subject that was left out. Each input sequence was compared with model for each of the six basic expressions. A dissimilarity score (Eq. (11)) was computed to measure the dissimilarity degree for each comparison. Finally, the minimal score was selected and the corresponding expression type indicated the expression of the input data. In this test, the subclass number c was set to be 4 and 20% pixels of each

Table 2
Summary of the system parameters (# stands for the short form of number)

Symbol	Description	Appeared in Eq.	Value
M	# of frames in each sequence	(1)	11
N_d	# of components used for classification	(2)	4
m	Fuzzy exponent	(3, 4, 5)	2
c	# of classification subclasses	(3, 4, 5)	4
col	# of pixels in each column	(9)	220
row	# of pixels in each row	(9)	200

Table 3
The recognition results

Label	Input					
	Smile	Surprise	Anger	Disgust	Fear	Sadness
Smile	89	1	1	2	1	7
Surprise	0	79	0	1	0	3
Anger	0	3	23	0	3	3
Disgust	1	0	0	28	0	1
Fear	1	2	1	1	49	2
Sadness	2	3	0	1	1	56
Accuracy (%)	95.7	89.8	92.0	84.8	90.7	77.8
Average (%)			88.8			

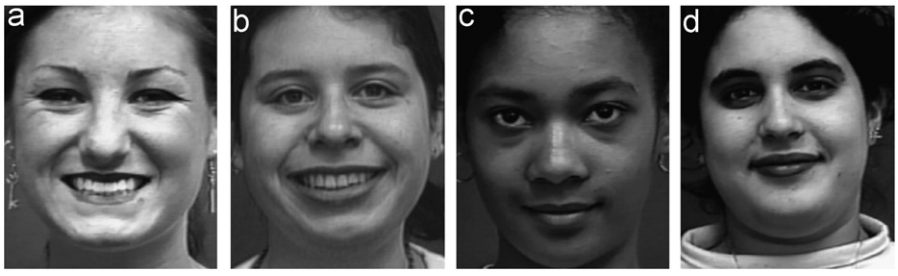

Fig. 7. Different degrees of mouth movements for the 'Smile' expression.

sequence were selected for classification since these were the optimal settings found in Section 4.2. The recognition results have been tabulated in Table 3.

From Table 3, it can be observed that the expression 'Smile' has achieved the best recognition rates. This can be explained as "when people smile, their mouths always moves more dramatically than when they perform others expressions". This makes the 'Smile' expression more distinctive. For 'Disgust' and 'Sadness', the facial muscle movements can be similar to other expressions. Even we get confused and have difficulty to identify these two expressions for some sequences. This could be the reason to explain the worse performance of these two expressions. In addition, another source that contributed additional errors could be due to the fact that different subjects had played the same expression in different manners and degrees. Four examples of the 'Smile' expressions with different degrees of maximum muscle distortion can be found in Fig. 7. This unpredicted variation had made the identification more difficult and complicate.

4.2. Empirical optimal setting

In previous section, we have discussed the recognition results with respect to the optimal setting of c as 4 and $N_{\%}$ being the number of the top 20% pixels that showed the highest expression activity. We have empirically varied these two parameters to search for the optimal setting for the system. The recognition rates with respect to changes of these two parameters are tabulated in Table 4.The optimal recognition rate is found to be 88.8% when c equals 4 and $N_{\%}$ equals 20%.

4.3. The effects of resolution and number of frames

The proposed system can be affected by two additional parameters, i.e. the image resolution and the number of frames per sequence. One can speculate that the performance should improve if higher resolution is used and more frames per expression sequence can be supplied since this is equivalent to supplying more and better information to the proposed system. On the other hand, one can also speculate whether it is possible to use lower resolution and less number of frames per sequence to simplify the computation process and at the same time maintain similar recognition performance without much degradation. We have performed experiments to measure the recognition rate with respect to these two parameters. The results are displayed in Fig. 8. Fig. 8(a) records different recognition rates according to different resolutions on the X-axis. The performance degraded gradually, and the system was able to give a recognition rate close to 0.57 even when the width was only 33 pixels. Fig. 8(b) displays the results with respect to different numbers of frames per sequence. The performance degraded slowly, and the system was able to give a recognition rate close to 0.6 even when the number of frames was cut down to 4. These experiments show that a working system can

Table 4
Recognition rates with respect to c and $N_{\%}$

Recognition rates		Selected pixel percentage ($N_{\%}$)					
		5%	10%	20%	30%	40%	50%
Subclass number (c)	2	81.1%	83.5%	87.9%	84.0%	83.2%	84.1%
	3	82.7%	84.6%	87.9%	82.7%	83.8%	84.6%
	4	82.4%	86.0%	88.8%	85.5%	85.7%	86.6%
	5	82.2%	84.6%	86.6%	83.5%	85.2%	86.0%
	6	81.1%	85.2%	85.8%	84.6%	84.4%	85.2%

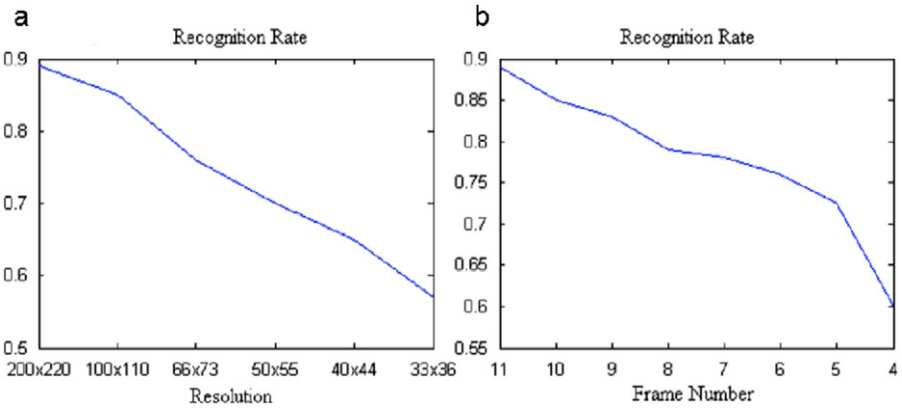

Fig. 8. Recognition rate with respect to (a) different resolutions and (b) different numbers of frames.

be built with lower resolution and smaller number of frames per sequence.

4.4. Comparison with others results

It is difficult to compare directly our results with other since different research groups had conducted different types of tests using different sets of data. All past experiments can be classified into two main categories approximately: using a static image or using a sequence of images (displaying one expression). Since we have mentioned previously that independent tests are more difficult in Section 4.1, we discuss and compare here only the independent test results. The recent recognition results (from 2002 to 2006) of these two types are listed in Tables 5 and 6 ,respectively. It is found that high recognition rates can be achieved when the numbers of human subjects, expression types and test sequences are small.

Using a static image: In Ref. [18], 70 unknown images of 10 persons were tested for recognition of seven expressions, i.e. the neutral expression was also tested. A mean accuracy rate of 83.3% was reported. Ma and Khorasani [29] reported a recognition result of 93.8% on a database of 60 men, each having five face images with the expressions neutral, smile, anger, sadness and surprise. Forty men were used for training a constructive NN and 20 men were used for testing. The neutral expression was not tested (recognized). Guo and Dyer [36] employed 213 images of seven expressions from 10 Japanese women as experiment data and reported an accuracy of 91.0%. In Ref. [35], Zheng et al. reported 77.1% and 79.2% accuracies on the JAFFE database and EKMAN database, respectively. Wu et al. in Ref. [37] reported an accuracy of 83.2% by classifying 183 frames from 10 subjects into six expression types. Our work has produced a recognition rate of 88.8% which is slightly lower than the results from Refs. [29,36]. However, they had used much less numbers of human subjects and tests. We employed 95 subjects and conducted 365 tests. In addition, Ref. [29] tested only four expression types. Hence, our result is better and more reliable.

Using a sequence of images: Kobayashi et al. [10] experimented with one human subject performing three expressions in 54 sequences. They achieved a recognition rate of 98.1%. In Ref. [13], 600 frames of one human subject were used for testing and a 96.8% recognition rate was obtained for six standard types of expressions. The 600 frames can be grouped into 60 expressions with each expression spanning about 10 frames. The work reported in Refs. [5,27] tested their system performances on the Cohn–Kanade database which is also being used in this study. They reported recognition rates of 84% and 81.8%, respectively. Our work has produced a recognition rate of 88.8% which is lower than the results from Refs. [10,13]. However, they tested their work using only one human subject that is far less desirable. Hence, our result is the best and more reliable.

In short, we had tested our work on a more difficult expression database with 95 subjects and 365 expression sequences. Our recognition rate of 88.8% has outperformed other who had also tested their systems on the same database.

Table 5
Results of recent work using static image

Approaches	# of human subjects	Expressions types	# of test (frames)	Recognition rates (%)
Abboud and Davoine [18]	10	7	70	83.3
Ma and Khorasani [29]	60	4	80	93.8
Zheng et al. [35]	–	6	183	77.1
Zheng et al. [35]	14	6	96	79.2
Guo and Dyer [36]	10	7	213	91.0
Wu et al. [37]	10	6	183	83.2

Table 6
Comparisons of recent work that use image sequences

Approaches	# of human subjects	Expression types	# of test frames or sequences	Recognition rates (%)
Kobayashi et al. [10]	1	3	54 sequences	98.1
Zhang and Ji [13]	1	6	600 frames	96.8
Pardàs et al. [5]	90	6	276 sequences	84
Cohen et al. [27]	53	6	318 sequences	81.8
The proposed FST	95	6	365 sequences	88.8

5. Conclusion

A novel efficient facial expression recognition system, namely fuzzy spatio-temporal (FST) modeling, has been developed in this work to analyze/recognize dynamic facial expressions. A sequence of images (displaying one expression) instead of just one image is analyzed. A novel expression descriptor has been derived to extract local temporal information at each pixel location using Fourier transform. A FCM (fuzzy C means) computation is then employed to model the temporal changes, and the extracted information is integrated with the spatial location to form expression models. Finally, an input is matched to the models using Hausdorff distance that tolerates misalignment and variations of facial component shape. The proposed system has been fine tuned and experimented with many sequences from the commonly used Cohn–Kanade facial expression database. The proposed approach has been shown to be effective, efficient and practical, with superior performance in comparison to other facial expression recognition systems. Future work will focus on employing 2D Fourier transform to get more accurate result to build intelligent and interactive system for human computer communication.

References

[1] P.S. Aleksic, A.K. Katsaggelos, Automatic facial expression recognition using facial animation parameters and multiStream HMMs, IEEE Trans. Inf. Forensics Secur. 1 (1) (2006) 3–11.
[2] F. Bourel, C. Chibelushi, A. Low, Recognition of facial expressions in the presence of occlusion, in: Proceedings of the 12th BMVC, vol. 1, Manchester, September 2001, pp. 213–222.
[3] M. Suwa, N. Sugie, K. Fujimora, A preliminary note on pattern recognition of human emotional expression, in: Proceedings of the 4th International Joint Conference on Pattern Recognition, 1978, pp. 408–410.
[4] I. Cohen, Facial expression recognition from video sequences: temporal and static modeling, CVIU Special Issue on Face Recognition, vol. 91, 2003, pp. 160–187.
[5] M. Pardàs, A. Bonafonte, J. Landabaso, Emotion recognition based on MPEG4 facial animation parameters, in: Proceedings of IEEE ICASSP, May 2002, pp. 3624–3627.
[6] Y. Yacoob, L.S. Davis, Recognizing human facial expressions from long image sequences using optical flow, IEEE Trans. Pattern Anal. Mach. Intell. 18 (6) (1996) 636–642.
[7] P. Ekman, W. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, Consulting Psychologists Press, Palo Alto, CA, 1978.
[8] I. Buciu, C. Kotropoulos, I. Pitas, ICA and Gabor representation for facial expression recognition, in: Proceedings of the 2003 International Conference on ICIP, vol. 3, September 2003, pp. 855–858.
[9] B. Abboud, F. Davoine, Bilinear factorisation for facial expression analysis and synthesis, IEE Proc. Vision Image Signal Process. 152 (3) (2005) 327–333.
[10] T. Kobayashi, Y. Ogawa, K. Kato, K. Yamamoto, Learning system of human facial expression for a family robot, in: Proceedings of 16th IEEE International Conference on Automatic Face and Gesture Recognition, May 2004, pp. 481–486.
[11] I. Cohen, N. Sebe, A. Garg, M.S. Lew, T.S. Huang, Facial expression recognition from video sequences, in: Proceedings of the 2002 ICME, August 2002, pp. 121–124.
[12] Y. Tian, T. Kanade, J. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115.
[13] Y. Zhang, Q. Ji, Active and dynamic information fusion for facial expression understanding from image sequences, IEEE Trans. Pattern Anal. Mach. Intell. 27 (5) (2005) 699–714.
[14] S. Ioannou, M. Wallace, K. Karpouzis, A. Raouzaiou, S. Kollias, Confidence-based fusion of multiple feature cues for facial expression recognition, in: Proceedings of 14th IEEE International Conference on Fuzzy Systems, 2005, pp. 207–212.
[15] N.B. Karayiannis, J.C. Bezdek, An intergrated approach to fuzzy learning vector quantization and fuzzy c-means clustering, IEEE Trans. Fuzzy Systems 5 (4) (1997) 622–628.
[16] J.N. Bassili, Emotion recognition: the role of facial movement and the relative importance of upper and lower area of the face, Pers. Soc. Psychol. 37 (1979) 2049–2059.
[17] G. Berks, D.G. Keyserlingk, J. Jantzen, M. Dotoli, H. Axer, Fuzzy clustering—a versatile mean to explore medical database, ESIT2000, Aachen, Germany.
[18] B. Abboud, F. Davoine, Appearance factorization based facial expression recognition and synthesis, in: Proceedings of 17th International Conference on Pattern Recognition, vol. 4, August 2004, pp. 163–166.

[19] C.C. Chibelushi, F. Bourel, Hierarchical multistream recognition of facial expressions, IEE Proc. Vision Image Signal Process. 151 (4) (2004) 307–313.
[20] B. Fasela, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognition 36 (2003) 259–275.
[21] I. Buciu, I. Pitas, Application of non-negative and local non negative matrix factorization to facial expression recognition, in: Proceedings of 17th International Conference on Pattern Recognition, vol. 1, August 2004, pp. 288–291.
[22] I. Cohen, F.G. Cozman, N. Sebe, M.C. Cirelo, T.S. Huang, Semisupervised learning of classifiers: theory, algorithms, and their application to human–computer interaction, IEEE Trans. Pattern Anal. Mach. Intell. 26 (12) (2004) 1553–1566.
[23] Y. Wang, H. Ai, B. Wu, C. Huang, Real time facial expression recognition with AdaBoost, in: Proceedings of 17th International Conference on Pattern Recognition, vol. 3, August 2004, pp. 926–929.
[24] S. Krinidis, I. Buciu, I. Pitas, Facial expression analysis and synthesis: a survey, in: Proceedings of 10th HCI International Conference on Human–Computer Interaction, 2003, pp. 1432–1433.
[25] J.J. Lien, T. Kanade, J.F. Cohn, C. Li, Detection, tracking, and classification of action units in facial expression, J. Robot Autonomous Systems 31 (1997) 131–146.
[26] N. Oliver, A. Pentland, F. Berard, LAFTER: a real-time face and lips tracker with facial expression recognition, Pattern Recognition 33 (7) (2000) 1369–1382.
[27] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, T. Huang, Learning Bayesian network classifiers for facial expression recognition using both labeled and unlabeled data, in: Proceedings of the 2003 IEEE CVPR, 2003.
[28] I. Essa, A. Pentland, Coding, analysis, interpretation, and recognition of facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 757–763.
[29] L. Ma, K. Khorasani, Facial expression recognition using constructive feedforward neural networks, IEEE Trans. Syst. Man Cybern. Part B 34 (3) (2004) 1588–1595.
[30] T. Kanade, J.F. Cohn, Y. Tian, Comprehensive database for facial expression analysis, in: Proceedings of 4th IEEE International Conference on Automatic Face and Gestures Recognition, 2000, pp. 46–53.
[31] R. Avent, C. Ng, J. Neal, Machine vision recognition of facial affect using back propagation neural networks, in: Proceedings of the 1994 IEEE Annual International Conference on Expression Recognition, 1994, pp. 1364–1365.
[32] C. Lisetti, D. Rumelhart, Facial expression recognition using a neural network, in: Proceedings of the 1998 International Conference on Flairs, 1998.
[33] R.Q. Feitosa, B.R. Vellasco, D.T. Oliveira, D.V. Andrade, S.A.R.S. Maffra, Facial expression classification using RBF and back-propagation neural networks, International Conference on ISAS, August 2000, pp. 73–77.
[34] L. Franco, A. Treves, A neural network face expression recognition system using unsupervised local processing, in: Proceedings of the 2001 International Symposium on Image and Signal Processing and Analysis, Croatia, 2001, pp. 628–632.
[35] W. Zheng, X. Zhou, C. Zou, C. Zhao, Facial expression recognition using Kernel Canonical Correlation Analysis (KCCA), IEEE Trans. Neural Networks 17 (1) (2006) 233–238.
[36] G. Guo, C. Dyer, Learning from examples in the small sample case: face expression recognition Systems, IEEE Trans. Man Cybernet. Part B 35 (3) (2005) 477–488.
[37] Y. Wu, H. Liu, H. Zha, Modeling facial expression space for recognition, in: Proceedings of International Conference on Intelligent Robots and Systems, August 2005, pp. 1968–1973.
[38] N. Morrison, Introduction to Fourier Analysis, Wiley-Interscience, New York, 1994.
[39] T. Xiang, M.K.H. Leung, Y. Chen, Facial morphing expression recognition, in: Proceedings of International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, March 2005, pp. 1–8.

About the Author—TUOWEN XIANG currently is a PhD candidate in School of Computer Engineering at Nanyang Technological University, Singapore. He received his bachelor degree from Zhejiang University, China, in 2002. His main area of research is human facial expression recognition.

About the Author—MAYLOR K.H. LEUNG received the BSc degree in physics from the National Taiwan University in 1979, and the BSc, MSc and PhD degrees in computer science from the University of Saskatchewan, Canada, in 1983, 1985 and 1992, respectively. Currently, Dr. Leung is an Associate Professor with Nanyang Technological University, Singapore. His research interest is in the area of computer vision, pattern recognition and image processing. Particular interest is on improving security using visual information. Some demo works on video surveillance, face recognition and shape analysis can be found in www.ntu.edu.sg/home/asmkleung. Area of interest: computer vision, pattern recognition and image processing. Particular interests are on line pattern analysis (e.g. Hausdorff distances of line and curve), video surveillance for abnormal human behavior detection, facial expression recognition, and computer aids for the visually impaired.

About the Author—SIU YEUNG CHO is an Assistant Professor in School of Computer Engineering at Nanyang Technological University (NTU), Singapore. Concurrently, he is the Deputy Director of Forensics and Security Laboratory (ForSe Lab) in the same school. Before joining NTU in 2003, he was a Research Fellow for The Hong Kong Polytechnic University and City University of Hong Kong between 2000 and 2003, where he worked some projects for neural networks and adaptive image processing. He also leaded a project of content-based image retrieval by novel machine learning model which is one of the projects attached in the Centre for Multimedia Signal Processing at PolyU HK. Dr. Cho received his PhD and BEng(Hons) from the City University of Hong Kong and University of Brighton, UK, respectively, all in electronic and computer engineering. His research interests include neural networks and its applications, image analysis and 3D computer vision. Dr. Cho has published more 40 technical papers. Dr. Cho is a member of IEE and IEEE.